Face Flow

Patrick Snape Anastasios Roussos Yannis Panagakis Stefanos Zafeiriou
Department of Computing, Imperial College London
180 Queens Gate, SW7 2AZ, London, U.K
{p.snape,troussos,i.panagakis,s.zafeiriou}@imperial.ac.uk

Abstract

In this paper, we propose a method for the robust and efficient computation of multi-frame optical flow in an expressive sequence of facial images. We formulate a novel energy minimisation problem for establishing dense correspondences between a neutral template and every frame of a sequence. We exploit the highly correlated nature of human expressions by representing dense facial motion using a deformation basis. Furthermore, we exploit the even higher correlation between deformations in a given input sequence by imposing a low-rank prior on the coefficients of the deformation basis, yielding temporally consistent optical flow. Our proposed model-based formulation, in conjunction with the inverse compositional strategy and low-rank matrix optimisation that we adopt, leads to a highly efficient algorithm for calculating facial flow. As experimental evaluation, we show quantitative experiments on a challenging novel benchmark of face sequences, with dense ground truth optical flow provided by motion capture data. We also provide qualitative results on a real sequence displaying fast motion and occlusions. Extensive quantitative and qualitative comparisons demonstrate that the proposed method outperforms state-of-the-art optical flow and dense non-rigid registration techniques, whilst running an order of magnitude faster.

1. Introduction

Computing optical flow in the presence of non-rigid deformations is an important and challenging task. It plays a significant role in a wide variety of problems such as medical imaging, dense non-rigid 3D reconstruction, dense 3D mesh registration, motion segmentation, video re-texturing and super-resolution. Broadly, optical flow methods describe procedures to relate pixels in one image to pixels in another image of the same object. They establish a displacement field that can be thought of as a sampling, or warping, of the input image back onto the reference image. Traditionally, optical flow is applied on a pair of consecutive frames of a sequence, treating one of the frames as the template. However, in terms of revealing the dynamics of a non-rigid scene, it is much more useful to estimate the optical flow between every frame of a long sequence and a common template. In this scenario, long-term dense 2D tracks across the sequence are established. We tackle the problem of multi-frame optical flow by focusing on a specific deformable object, the human face.

The most straight-forward way to estimate a multi-frame optical flow is to apply an algorithm that solves the traditional two-frame optical flow problem between every frame and the template independently. However, the fact that we have to deal with long sequences poses major difficulties. For example, the point displacements between the template and a frame can be substantially large and severe occlusions of parts of the template can occur in some frames. Even state-of-the-art two-frame optical flow methods that are especially designed to deal with large displacements or occlusions, *e.g.* [9, 27], are prone to fail. This is because they lack any additional cues that could help them estimate an appropriate initialisation or to disambiguate severe occlusions.

An alternative solution to the multi-frame optical flow problem, also based on two-frame optical flow, is to estimate flow between consecutive frames and then combine the various solutions. A simple integration of the solutions to obtain long-term 2D tracks is prone to drift due to error accumulation [11, 9]. This can be improved by the automatic detection of occlusions, gross errors, and other ambiguities [32, 34, 31, 29, 25], but any such solution is still limited by the accuracy of the initial two-frame optical flow estimations that are completely local in time and do not exploit any temporal cues.

Several recent methods solve the multi-frame optical flow problem directly, by implicitly taking into account the rich temporal information that is present in non-rigid scenes [15, 36, 35, 28, 30, 14]. For example, the long-term 2D trajectories of points on a surface undergoing non-rigid deformation are highly correlated and can be compactly described via a linear combination of a low-rank trajectory

basis. This basis is typically learnt from the input sequence itself. In this way, these methods are more robust to occlusions and yield a temporally coherent result. However, they rely only on some generic spatial and temporal regularisation priors, applicable to any deformable object and do not utilise any prior knowledge about the specific object observed in the scene. This makes them fail in more challenging conditions that often occur in real-world scenes, such as severe occlusions or significant illumination changes, which cannot be disambiguated by temporal regularisation alone. Furthermore, the memory and runtime efficiency of all existing multi-frame optical flow methods is limited by the fact that they have to estimate a very large number of parameters, i.e. a set of parameters for every pixel of the template. Specially designed parallelisable algorithms, such as primal-dual optimisation schemes [37, 14] can be adopted. However, they are only efficient on recent GPU hardware.

In this paper, we overcome the aforementioned limitations by incorporating a face-specific deformation model into the multi-frame optical flow estimation. We assume a learnt deformation basis, rather than one calculated directly from the sequence itself. We focus on human faces which are a very commonly considered object in computer vision, and dense face correspondences are required in many research areas and applications. That is, the establishment of dense correspondences of deformable faces is the first step towards high-performance facial expression recognition [18], facial motion capture [8] and 3D face reconstruction [13]. Nevertheless, computing dense face correspondences has received limited attention [12, 40]. This is attributed to the difficulty of developing a statistical model for dense facial flow due to the in-ability of humans to densely annotate sequences and the limited robustness of the optical flow techniques [12]. Hence, the research community has focused on developing statistical facial models built on a sparse set of landmarks [39], which provide limited accuracy to the recognition of subtle expressions [19]. In this paper, we build the first, to the best of our knowledge, statistical models of dense facial flow by capitalising on the success of recent optical flow techniques applied to densely tracking image sequences [14]. Due to the use of the statistical low-rank model, and in contrast to existing approaches, our method is able to deal with particularly challenging conditions such as severe occlusions of the face and strong illumination changes. Furthermore, the introduction of a known deformation basis drastically reduces the dimensionality of the multi-frame optical flow problem and leads to a very efficient algorithm.

2. Contributions

We formulate a novel energy minimisation problem for the robust estimation of multi-frame optical flow in an expressive sequence of facial images. Given that the range of motion expressible by a human face is limited, and that faces themselves are well known to be highly correlated and compactly described by a low-dimensional subspace, we build a dense deformation basis for faces.

Furthermore, we exploit the even higher correlation between face deformations in a specific input sequence by imposing a low-rank prior on the coefficients of the deformation basis. This acts as a data-specific regularisation term leading to temporally consistent optical flow. We also incorporate a sparse landmark prior term to guide the flow estimation in sparse point locations that are accurately predicted by a state-of-the-art face alignment method [16]. Finally, we formulate the photometric cost by utilising a state-of-the-art dense feature descriptor that offers robustness even with the usage of a simple quadratic penaliser. Our proposed model-based problem formulation, in conjunction with the inverse compositional strategy and low-rank matrix optimisation that we adopt, leads to a highly efficient algorithm for calculating optical flow across a facial sequence. For experimental evaluation, we show quantitative experiments on a very challenging novel benchmark of face sequences with dense ground truth optical flow based on motion capture data. We also provide qualitative results on a real sequence displaying fast motion and natural occlusions.

3. Further Related Work

There is a very large body of work on facial alignment that largely revolves around the concept of identifying a set of sparse target landmarks within an image. The most relevant algorithm to our proposed method is that of the Active Appearance Model (AAM) [10], particularly the variation by Baker and Matthews [24] that relates AAMs to the Lucas-Kanade [22, 7] optical flow literature. However, our method does not incorporate an appearance model and relies on a single given template image and is thus closer in nature to the original Lucas-Kanade algorithm. Our algorithm also places a low-rank constraint on the shape model coefficients enforcing a form of temporal consistency, which has not been previously considered.

It is also important to note that, other than a couple of examples [26, 4], the AAM literature has focused on the recovery of sparse landmarks, not a dense motion field as in this work. Even in the cases of [26, 4], the warping of the input images is achieved via an interpolation method such as piecewise affine or thin-plate splines. In this work, our warping method is derived directly from the deformation basis itself and thus recovers dense correspondences. As we show in Section 4, the linear nature of our warp allows us to derive a very efficient optimisation strategy, based on the Inverse Compositional algorithm proposed by Baker and Matthews [7].

Finally, the work of Kemelmacher-Schlizerman *et al.*

[17] is relevant as it considers learning deformation fields between images of faces. However, [17] considers an optical flow method as a key component of the method and does not propose a novel optical formulation, in contrast to our proposed model-based algorithm.

4. Face Flow Representation

Let us assume that the input face video contains F frames. Let $M \subset \mathbb{R}^2$ be a 2D domain that corresponds to the region of a mean face in neutral expression that will be used as a template. We seek to estimate a function $\boldsymbol{u}(\boldsymbol{x}; f) : M \times \{1, \dots, F\} \to \mathbb{R}^2$ that represents the optical flow from this domain to every frame of the input sequence. More precisely, this function establishes the correspondence between every facial point $\boldsymbol{x}$ in the domain M and its location at every frame index f, which is given by the warping function $W_f(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x}; f)$. This warping function registers the f-th frame to the domain M.

Exploiting prior knowledge about the warping functions that are yielded from facial deformations, we adopt a linear model for every $W_f(\boldsymbol{x})$ as follows:

$$W_f(\boldsymbol{x}) = W(\boldsymbol{x}; \boldsymbol{c}_f) = \langle \boldsymbol{B}(\boldsymbol{x}), \boldsymbol{c}_f \rangle, \, \boldsymbol{x} \in M, \qquad (1)$$

where $\boldsymbol{B} : M \to \mathbb{R}^D$ is a learnt basis of facial deformations that contains D basis vector elements and is common to all frames. Also, $\boldsymbol{c}_f \in \mathbb{R}^D$ is the coefficient vector for the f-th frame. $\boldsymbol{B}(\boldsymbol{x})$ is constructed a priori during a training process and therefore, for an input face video, we transform the multi-frame optical flow estimation to the estimation of the following $D \times F$ matrix:

$$\boldsymbol{C} = [\boldsymbol{c}_1 \cdots \boldsymbol{c}_f \cdots \boldsymbol{c}_F] \, , \qquad (2)$$

The f-th column of this matrix contains the coefficients that yield the warping function (and thus define the optical flow) for the f-th frame of the video. Following AAMs [10], the first 4 components of these coefficients, which correspond to the first 4 rows of the coefficient matrix $\boldsymbol{C}$, control the similarity transformation that rigidly-aligns the template to every frame. The remaining components control the non-rigid deformations. Therefore, we decompose $\boldsymbol{C}$ into the following sub-matrices:

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_s \\ \boldsymbol{C}_{nr} \end{bmatrix}, \qquad (3)$$

where $\boldsymbol{C}_s$ and $\boldsymbol{C}_{nr}$ correspond to the similarity and non-rigid part of the facial deformations respectively. $\boldsymbol{C}_s$ is a $4 \times F$ matrix and $\boldsymbol{C}_{nr}$ is a $K \times F$ matrix, where $K = D - 4$ is the rank of non-rigid deformations of the model.

5. Proposed Energy

Let $\boldsymbol{I}(\boldsymbol{x}; f) : \Omega \times \{1, \dots, F\} \to \mathbb{R}^{N_c}$ be the N_c-channel sequence of frames of the input video, where Ω is the rectangular image domain that corresponds to this video. The channels of the input frames originate from the application of some appropriate feature descriptor.

As a preprocessing step, a frame of the sequence which is as close as possible to a frontal pose and a neutral expression is selected as reference. This frame is warped to the template domain M in order to match a mean face. This selection and warping estimation can be easily done automatically by fitting the face with an automatic facial alignment method [16, 24]. In this case, we assume that there is a known correspondence between the sparse points found by the alignment method and the reference frame of our basis. Once the sparse landmarks have been acquired, it is simple to warp the image into the reference frame using a warping function such as piecewise affine. This procedure is identical to the one performed when building Active Appearance Models, with the exception of only being performed on a single template image. In short, this warped reference frame defines the template image $\boldsymbol{T}(\boldsymbol{x}) : M \to \mathbb{R}^{N_c}$.

We also consider the case where, as further preprocessing, a sparse set of facial landmarks is localised and tracked in the video. Let L be the total number of landmarks and $\boldsymbol{\ell}_{i,f} \in \mathbb{R}^2$ the position of the i-th landmark on the f-th frame. In addition, let $\hat{\boldsymbol{\ell}}_i \in \mathbb{R}^2$ be the position of the i-th landmark on the template image, which is computed by applying the warping function on the corresponding landmark of the reference frame.

We propose to estimate the face flow through the minimisation of the following energy:

$$E(\boldsymbol{C}) = E_{img}(\boldsymbol{C}) + \beta E_{land}(\boldsymbol{C}) \, , \qquad (4)$$

under the low-rank constraint:

$$\text{rank}(\boldsymbol{C}_{nr}) \leq \lambda, \qquad (5)$$

where $\boldsymbol{C}_{nr}$ is the non-rigid part of $\boldsymbol{C}$ (3). E_{img} is an image data term and E_{land} is a landmark term. The positive weight β controls the balance between these terms, whereas the integer $0 \leq \lambda \leq K$ is the imposed maximum rank of non-rigid deformations for the input sequence. We now define and explain the different parts of this minimisation problem.

The first term (E_{img}) enforces consistency of the feature descriptor values of every point of the template over all frames:

$$E_{img} = \sum_{f=1}^{F} \int_M \| \boldsymbol{T}(\boldsymbol{x}) - \boldsymbol{I}\left(W(\boldsymbol{x}; \boldsymbol{c}_f) \, ; \, f\right) \|^2 \, \mathrm{d}\boldsymbol{x}, \qquad (6)$$

In general, such an image data term could be grossly affected by artifacts in the image, such as illumination variation and external occlusions. Therefore, it is common to use a robust penaliser rather than the quadratic term shown in (6). However, we act on recent advancements in facial alignment algorithms [5] that suggest that densely sampled

feature descriptors can vastly improve the performance of alignment algorithms without sacrificing the efficiency of a quadratic optimisation. The use of dense descriptors is similar to SIFTFlow [20], where SIFT [21] features are densely sampled at every pixel in order to improve optical flow.

The second term (E_{land}) is a quadratic prior that ensures that the estimated face flow is in accordance with the landmark information on the corresponding sparse points in every frame:

$$E_{land} = \sum_{f=1}^{F} \sum_{i=1}^{L} \left\| W(\hat{\boldsymbol{\ell}}_i; \boldsymbol{c}_f) - \boldsymbol{\ell}_{i,f} \right\|^2, \tag{7}$$

Regarding the low-rank constraint (5), it is natural to assume that the deformations of the face over time are highly correlated and thus lie in a low-dimensional subspace. However, the similarity transformations describing the face motion are, in general, not sufficiently correlated with the non-rigid deformations that the face undergoes. For example, the similarity transformations often originate from camera motion. Consequently, we penalise the number of independent components needed to describe the non-rigid face deformations of the specific input sequence and thus impose $\boldsymbol{C}_{nr}$ to be of low-rank.

5.1. Estimation of the Sparse Landmarks

In order to estimate sparse landmarks, we can make use of state-of-the-art, extremely efficient facial alignment methods [16, 6, 1, 2]. State-of-the-art methods, such as that by Kazemi *et al*. [16], execute in under a millisecond, and can provide landmark localisation errors of within 3 pixels on average for extremely challenging unconstrained images. In this paper, when we consider estimating landmarks, we use the method of [16] in conjunction with a robust face detector [42].

5.2. Learning the Deformation Basis

Learning the deformation basis is a very challenging issue and most likely the reason why there is little research in building dense facial deformation models. However, inspired by the performance of recent optical methods, such as that of Garg *et al*. [14], we chose to build our basis using the output of optical flow methods. To realise this, we chose the optical flow method of Garg *et al*. [14], augmented with an additional quadratic penalty involving automatically estimated sparse landmarks. This additional penalty was found to significantly improve the performance of [14] in expressive sequences, such as wide openings of the mouth.

We propose to learn a set of trajectories over a number of sequences, each with a differing reference frame. We are thus faced with the problem of achieving correspondence between these reference frames for the construction of the deformation basis. Given that each frame contains estimated sparse landmarks, we calculate the mean position of each landmark and define the area of spatial support for our deformation basis, M, to be the pixels that are situated inside the convex hull of the these positions. Once the reference frame is constructed, each set of trajectories is converted into endpoints for each frame, analogous to dense landmarks for the image, and sampled into the reference frame using a thin-plate splines warp parametrised by the automatically estimated landmarks. Finally, given that we have a set of dense landmarks in correspondence, we perform a Procrustes alignment in order to normalise any scale issues that may be present.

6. Optimisation of the Proposed Energy

The image data term is highly non-convex, therefore we have to adopt an iterative linearisation scheme. For this scheme to be computationally efficient, we consider an inverse compositional (IC) strategy. In every iteration, we seek to update the current estimate $\tilde{\boldsymbol{C}} = [\tilde{\boldsymbol{c}}_1 \cdots \tilde{\boldsymbol{c}}_F]$ of the coefficient matrix. In addition, we consider a spatial discretisation of E_{img} on a regular pixel grid with unary steps. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_P$ be the 2D locations of the P pixels that lie within the domain M.

The IC algorithm, as proposed by [7], is a very efficient method of solving a parametrised image alignment problem, which corresponds to minimising solely the image data term E_{img} of the proposed energy for every frame of the sequence. Given a single template image $\boldsymbol{T}$ and a single input image $\boldsymbol{I}$, the classical Lucas-Kanade [22] problem is given by

$$\sum_{p=1}^{P} \left\| \boldsymbol{T}(W(\boldsymbol{x}_p; \Delta\boldsymbol{c})) - \boldsymbol{I}\left(W(\boldsymbol{x}_p; \boldsymbol{c})\right) \right\|^2, \tag{8}$$

which is minimised for $\Delta\boldsymbol{c}$, the parameters of the warp for a single image. Here, $\boldsymbol{T}(W(\boldsymbol{x}_p; \Delta\boldsymbol{c}))$ denotes the template warped around the current linearised estimate of $\Delta\boldsymbol{c}$. The IC algorithm is so efficient because we assume that we linearise (8) around $\Delta\boldsymbol{c}$ and thus the template is fixed and does not require warping during the updates. To update the parameters $\boldsymbol{c}$, a compositional update is performed: $W(\boldsymbol{x}_p; \Delta\boldsymbol{c}) \leftarrow W(\boldsymbol{x}_p; \Delta\boldsymbol{c}) \cdot W(\boldsymbol{x}_p; \Delta\boldsymbol{c})^{-1}$. This update ensures that the derivative with respect to the warp is also fixed and therefore we arrive at the extremely efficient update for $\boldsymbol{c}$:

$$\Delta\boldsymbol{c} = (\boldsymbol{J}^T\boldsymbol{J})^{-1} \sum_{p=1}^{P} \boldsymbol{J}^T[\boldsymbol{I}\left(W(\boldsymbol{x}_p; \boldsymbol{c})\right) - \boldsymbol{T}(\boldsymbol{x}_p)], \tag{9}$$

where $\boldsymbol{J} = \nabla\boldsymbol{T}\frac{\partial W}{\partial \boldsymbol{c}}$, and the derivative $\frac{\partial W}{\partial \boldsymbol{c}}$ is taken around $(\boldsymbol{x}_p; \boldsymbol{0}) = \boldsymbol{B}(\boldsymbol{x}_p)$. Therefore, the entire Hessian term $\boldsymbol{H} = \boldsymbol{J}^T\boldsymbol{J}$, does not depend on $\boldsymbol{c}$ and can be precomputed. Unlike in most previous works in the area, our motion model

is completely translational and thus does not involve a complicated compositional update. In fact, it can be shown that our compositional update has the form $\boldsymbol{c} \leftarrow \boldsymbol{c} - \Delta\boldsymbol{c}$ and is thus equivalent to the additive parameter update scheme [3].

Returning to the optimisation of the proposed energy, the image data term can be approximated as (after the IC strategy and the spatial discretisation):

$$E_{img} \approx \sum_{f=1}^{F} \sum_{p=1}^{P} \left\| \boldsymbol{T}(W(\boldsymbol{x}_p; \Delta\boldsymbol{c}_f)) - \boldsymbol{I}\left(W(\boldsymbol{x}_p; \tilde{\boldsymbol{c}}_f)\,;\, f\right) \right\|^2, \tag{10}$$

where $\Delta\boldsymbol{c}_f$ are the additive warp parameters for frame f. Note that $\Delta\boldsymbol{c}_f = \boldsymbol{c}_f - \tilde{\boldsymbol{c}}_f$, a relation that we use since in our formulation, in contrast to the traditional IC algorithm, we incorporate terms that depend directly on $\boldsymbol{c}_f$. By considering linearisations of the template in the above equation and rewriting the terms using compact matrix notation over all pixels and frames, the total proposed energy becomes:

$$E(\boldsymbol{C}) \approx \left\| \boldsymbol{R} + \boldsymbol{J}(\boldsymbol{C} - \tilde{\boldsymbol{C}}) \right\|_F^2 + \beta \left\| \boldsymbol{B}_\ell \boldsymbol{C} - \boldsymbol{L}_{loc} \right\|_F^2, \tag{11}$$

where $\boldsymbol{R}$ is a $(PN_c) \times F$ matrix that contains the error residuals $\boldsymbol{T}(\boldsymbol{x}_p)) - \boldsymbol{I}\left(W(\boldsymbol{x}_p; \tilde{\boldsymbol{c}}_f)\,;\, f\right)$ for all pixels p and frames f. Also, $\boldsymbol{J}$ is a $(PN_c) \times D$ matrix that contains the template Jacobian $\nabla\boldsymbol{T}(\boldsymbol{x}_p)\boldsymbol{B}(\boldsymbol{x}_p)$ for all pixels. Finally, $\boldsymbol{B}_\ell$ is a $2L \times D$ matrix consisting of the deformation basis evaluated at the locations of the landmarks on the template and $\boldsymbol{L}_{loc}$ is a $2L \times F$ matrix with the coordinates of the landmarks in all frames:

$$\boldsymbol{B}_\ell = \begin{bmatrix} \boldsymbol{B}(\hat{\boldsymbol{\ell}}_1) \\ \vdots \\ \boldsymbol{B}(\hat{\boldsymbol{\ell}}_L) \end{bmatrix}, \boldsymbol{L}_{loc} = \begin{bmatrix} \boldsymbol{\ell}_{1,1} & \cdots & \boldsymbol{\ell}_{1,F} \\ \vdots & & \vdots \\ \boldsymbol{\ell}_{L,1} & \cdots & \boldsymbol{\ell}_{L,F} \end{bmatrix}, \tag{12}$$

Using the decomposition of $\boldsymbol{C}$ in a similarity and a non-rigid part (3), the energy (11) is written as:

$$\begin{aligned} &E(\boldsymbol{C}_s, \boldsymbol{C}_{nr}) \approx \\ &\left\| \boldsymbol{R} + \boldsymbol{J}_{nr}(\boldsymbol{C}_{nr} - \tilde{\boldsymbol{C}}_{nr}) + \boldsymbol{J}_s(\boldsymbol{C}_s - \tilde{\boldsymbol{C}}_s) \right\|_F^2 \\ &+\beta \left\| \boldsymbol{B}_{\ell nr}\boldsymbol{C}_{nr} - \boldsymbol{B}_{\ell s}\boldsymbol{C}_s - \boldsymbol{L}_{loc} \right\|_F^2, \end{aligned}$$

Consequently, we propose to solve the following rank constraint optimisation problem:

$$\min_{\boldsymbol{C}_s, \boldsymbol{C}_{nr}} E(\boldsymbol{C}_s, \boldsymbol{C}_{nr}) \quad \text{s.t.} \quad \text{rank}(\boldsymbol{C}_{nr}) \leq \lambda, \tag{13}$$

Although (13) is a non-convex problem, it can be solved efficiently by employing a block-coordinate descent (BCD) scheme. Let t be the iteration index. The iteration of BCD for (13) reads as follows:

$$\boldsymbol{C}_s[t+1] = \min_{\boldsymbol{C}_s[t]} E(\boldsymbol{C}_s[t], \boldsymbol{C}_{nr}[t]), \tag{14}$$

$$\begin{aligned} \boldsymbol{C}_{nr}[t+1] = &\min_{\boldsymbol{C}_{nr}[t+1]} E(\boldsymbol{C}_s[t+1], \boldsymbol{C}_{nr}[t]), \\ &\text{s.t.} \quad \text{rank}(\boldsymbol{C}_{nr}) \leq \lambda. \end{aligned} \tag{15}$$

The sub-problem (14) is a least-squares problem admitting a closed-form solution.

The sub-problem (15) is also solved in closed-form. First, let us define the matrices

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{R} + \boldsymbol{J}_s(\boldsymbol{C}_s - \tilde{\boldsymbol{C}}_s) \\ \boldsymbol{B}_{\ell s}\boldsymbol{C}_s - \boldsymbol{L}_{loc} \end{bmatrix}, \boldsymbol{Q} = \begin{bmatrix} \boldsymbol{J}_{nr} \\ \boldsymbol{B}_{\ell nr} \end{bmatrix}, \tag{16}$$

with $\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}$ being the Thin Singular Value Decomposition of $\boldsymbol{Q}$, $\boldsymbol{Q}^\dagger$ denoting the pseudo-inverse of $\boldsymbol{Q}$, $\boldsymbol{\Xi} = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{A}$, and $\boldsymbol{\Xi}_{(\lambda)}$ being the λ-rank approximation of $\boldsymbol{\Xi}$. Then by using (16), (15) is written as:

$$\min_{\boldsymbol{C}_{nr}} \left\| \boldsymbol{A} - \boldsymbol{Q}\boldsymbol{C}_{nr} \right\|_F^2 \quad \text{s.t.} \quad \text{rank}(\boldsymbol{C}_{nr}) \leq \lambda, \tag{17}$$

The closed form solution of (17) is given by [33]:

$$\boldsymbol{C}_{nr} = \boldsymbol{Q}^\dagger \boldsymbol{\Xi}_{(\lambda)}. \tag{18}$$

The convergence of the proposed BCD algorithm is guaranteed since the objective function is differentiable and involves two blocks of variables [23].

7. Experiments

In this section, we describe the set of qualitative and quantitative experiments that we conducted in order to demonstrate the effectiveness of our proposed algorithm, Face Flow. In order to verify that our method is competitive with the state-of-the-art, we compared against the methods of Garg *et al.* [14] (denoted MFSF), Revaud *et al.* [27] (denoted EPICFlow), Liu *et al.* [20] (denoted SIFTFlow) and the large displacement optical method of Brox *et al.* [9] (denoted LDOF). We also provide two formulations of our method, one which enforces the low-rank constraint on the coefficients and one which does not. The latter corresponds to the choice of $\lambda = k$. We denote these two methods Face Flow Low-Rank (LR) and Face Flow Full-Rank (FR). This self evaluation is particularly useful for demonstrating the importance and effectiveness of the low-rank constraint for multi-frame facial flow.

To effectively evaluate Face Flow, we propose a novel ground truth dataset formed from facial motion capture data [43]. The sequence that we evaluate must be in correspondence and ideally contain an interesting sequence of deformations. Performance capture data is ideal for this purpose, as it is necessarily in correspondence and often deals with

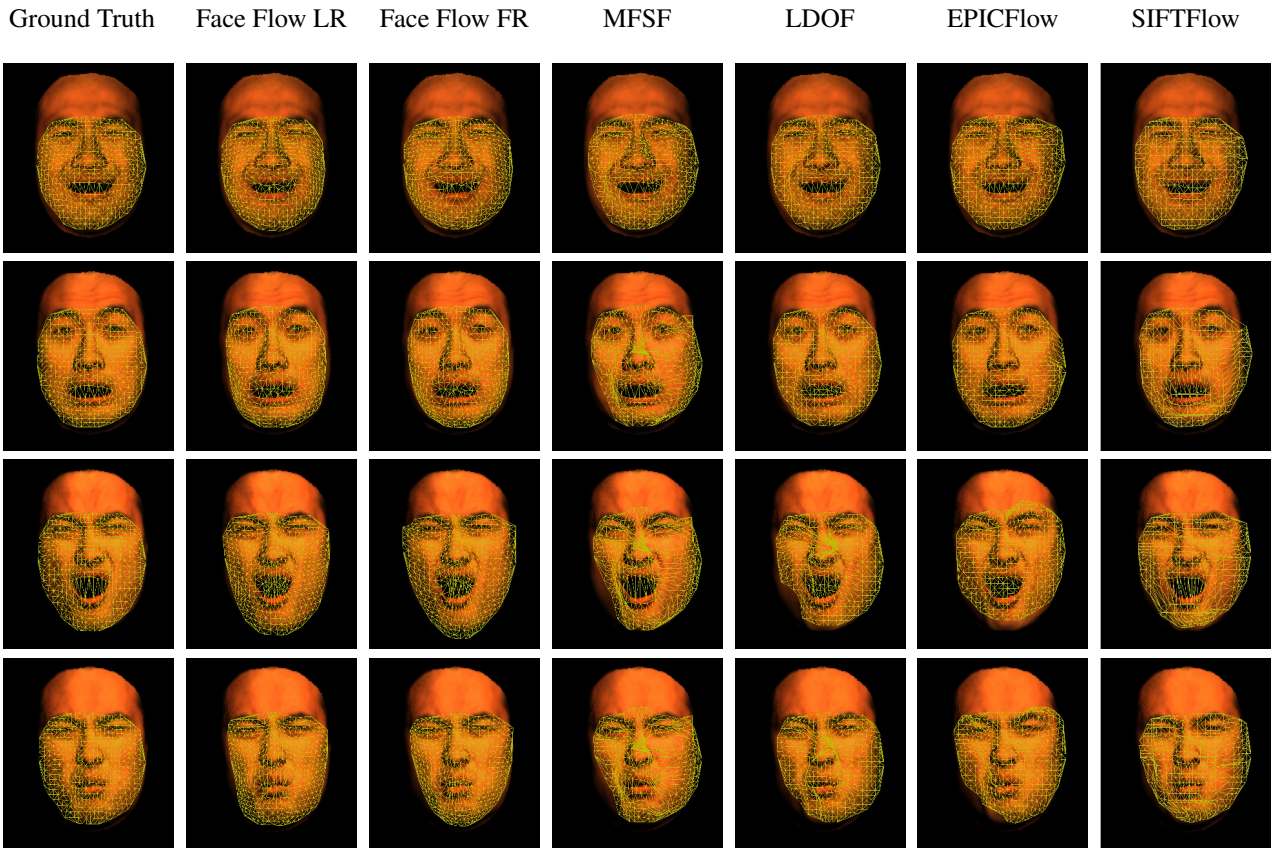

Figure 1: Example endpoint results for the synthetic sequence including illumination variation. Each row shows a different frame of the 280 frame synthetic sequence (Frame 16, 54, 139 and 233 from top to bottom)

actors portraying scripted, emotional content. We also evaluate face flow on a realistic sequence portraying a number of common issues in facial videos, including motion blur and occlusions. Please see the supplementary material for video examples of the sequences and further results.

7.1. Practical Deformation Basis Construction

Our facial deformation basis is built by applying MFSF multi-frame optical flow method [1] of Garg *et al*. [14] on the facial expression database BU4D [41]. We chose this database due to the large range of expression present and the fact that is captured at a high frame rate, which is ideal for the technique of [14]. However, in order to further improve the performance of [14], we augmented the energy to include an extra quadratic landmark constraint. This landmark constraint takes a similar form to the landmark constraint proposed in this paper, and was found to improve the results considerably in sequences that displayed particularly expressive emotion, such as surprise.

The BU4D database consists of 102 subjects displaying 6 canonical expression, from neutral to the apex of the emotion. We selected a neutral frame for every sequence and used this as the reference frame for the method of [14]. After computing trajectories for each sequence, we constructed a reference frame for our deformation basis using the mean of the neutral images we selected previously. We then applied principal component analysis (PCA) to learn the linear deformation model as described in Section 5.2. Experimentally, we found that $k = 20$ principal components of non-rigid deformation accounts for 95% of the variance. We note that the BU4D data shows frontal faces and thus our model does not capture out-of-plane rotation. However, there is no practical reason that our PCA basis could not capture this kind of deformation.

In order to improve the robustness of our algorithm, we adopt a pseudo coarse-to-fine strategy for basis construction and create three bases of increasing scale. This is also commonly employed within optical algorithms to improve robustness. In all of the following experiments, the feature descriptor employed is the dense SIFT feature.

7.2. Motion Capture Data

In this experiment, we use the performance capture dataset provided by Zhang *et al*. [43] to generate three

[1] Code publicly available at https://bitbucket.org/troussos/mfsf/

novel ground truth sequences consisting of 280 frames. This ground truth is provided by rendering the sequence of meshes in a fixed pose, which yields both a texture and a set of vertices in the scene. These vertices can then be used in order simply calculate flow for the face throughout the image.

As mentioned, we rendered three sets of texture, all with the same underlying geometry, to provide evidence of the robustness of Face Flow to challenging conditions. In the first sequence, we rendered unmodified textures. This is a baseline in order to show the performance of state-of-the-art methods for facial data. Since our basis is trained using the output from [14], we do not expect to outperform other optical flow techniques on this sequence. The second sequence was rendered by rendering a periodically moving light source around the face. This is challenging for the data term and helps to demonstrate the robustness of our chosen feature descriptor. The final sequence is highly challenging. It contains the periodic illumination variation from the previous sequence and also an artificial occlusion in the form of a smoothly translating hand.

In order to initialise our Face Flow algorithm, we manually annotated the first frame as the reference frame. To obtain an initial estimate of the coefficient matrix $\mathbf{C}$, the landmark constraint quadratic term is solved, which provides a reasonable estimate of the initial shape. Once solved in the reference frame, this initialisation was propagated across every frame in the sequence. The landmark constraint was otherwise not utilised in this experiment.

To evaluate the performance of the methods, we computed the root mean squared error of the endpoints (RMSE), shown in Table 2. As expected Face Flow does not outperform state-of-the-art methods in the original un-tampered sequence. However, in the more interesting case of the illumination variation and occlusion sequences, our Face Flow method with low-rank constraint (Face Flow LR), performs the best. We also note that the low-rank method of our technique significantly outperforms the full-rank version, particularly in the occluded sequence. An example set of endpoints is given in Figure 1. Note how the face deformation is well localised and unlikely to undergo any gross deformations due to being constrained by a statistical basis.

Finally, to provide evidence as to the stability of our algorithm, and the the effect of the low-rank constraint on the outcome of the sequence, we present Figure 2. This figure shows the mean per-frame endpoint error across the sequence. The Face Flow methods are given by the blue and green lines. Note how stable the Face Flow LR method is, particularly when compared to the Face Flow FR method. We believe this effectively demonstrates the positive effect enforcing soft temporal consistency can have.

FF LR	FF FR	MFSF	LDOF	EF [27]	SF [20]
1.4	0.5	20	15	50	15

Table 1: Run times *per frame (in seconds)* for the 150 frame sequence shown in Figure 3. Images are 640×480. Performed on an Intel Xeon E5-1650 3.20GHz (32GB RAM). All times are approximate and averaged over multiple runs. EPICFlow times are dominated by DEEPMatching [38] (48s).

	Original		Illum.		Ilum.+Occ.	
	RMSE	AE95	RMSE	AE95	RMSE	AE95
FF LR	2.95	5.52	**3.56**	**6.63**	**4.48**	**8.47**
FF FR	3.24	6.01	3.76	7.02	5.83	11.50
MFSF	1.73	3.20	6.33	13.68	8.25	17.30
LDOF	**1.56**	**2.79**	4.84	9.98	6.54	11.44
EF [27]	1.66	3.25	4.02	9.61	5.15	11.61
SF [20]	2.65	5.15	4.89	11.81	11.82	23.05

Table 2: RMSE and 95% average endpoint error for the synthetic data.

7.3. Real Sequence

For this experiment, we provide results on a real sequence consisting of 150 frames of a young woman watching a video. In this sequence, she reacts negatively and clearly presents discomfort by touching her face and shifting in her seat. This amounts to challenging variation in the sequence, including occlusions from her hand and motion blur from movement. In this sequence, we also provide evidence as to the benefit of incorporating the landmark constraint. The sequence was automatically landmarked using [16], then our method was initialised with these landmarks for every frame. We also used the landmarks for the quadratic penalizer term.

As Figure 3 shows, Face Flow LR performs very well in this sequence. In particular, we note that our method is very robust to the presence of occlusions, aided by the sparse landmarks provided by [16]. We also provide Table 1 which demonstrates that Face Flow is an order of magnitude more efficient than the other considered methods for this sequence.

8. Conclusion

We presented an optical flow method that incorporates a dense basis of facial shape. We evaluated our method on a ground truth motion capture sequence and demonstrated that our proposed algorithm, Face Flow, outperforms other state-of-the-art optical flow methods. In particular, the introduction of a low-rank constraint yields a robust multi-frame optical flow technique. As future work, we intend to investigate a more complex dataset that would enable Face

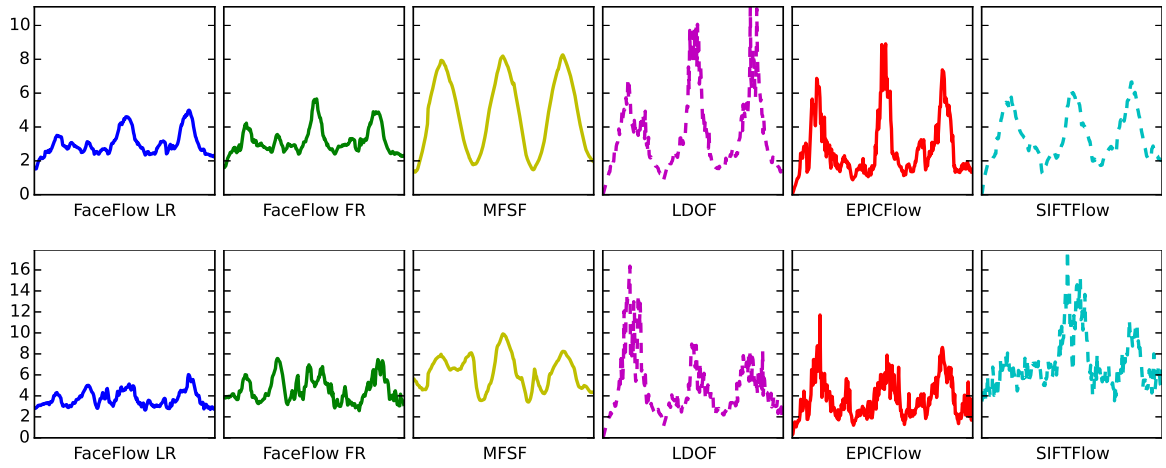

Figure 2: The average endpoint error calculated for each frame of the mocap sequence. Vertical axis is average endpoint error, horizontal is frame number. Top row is the illumination sequence, bottom row is illumination + occlusion.

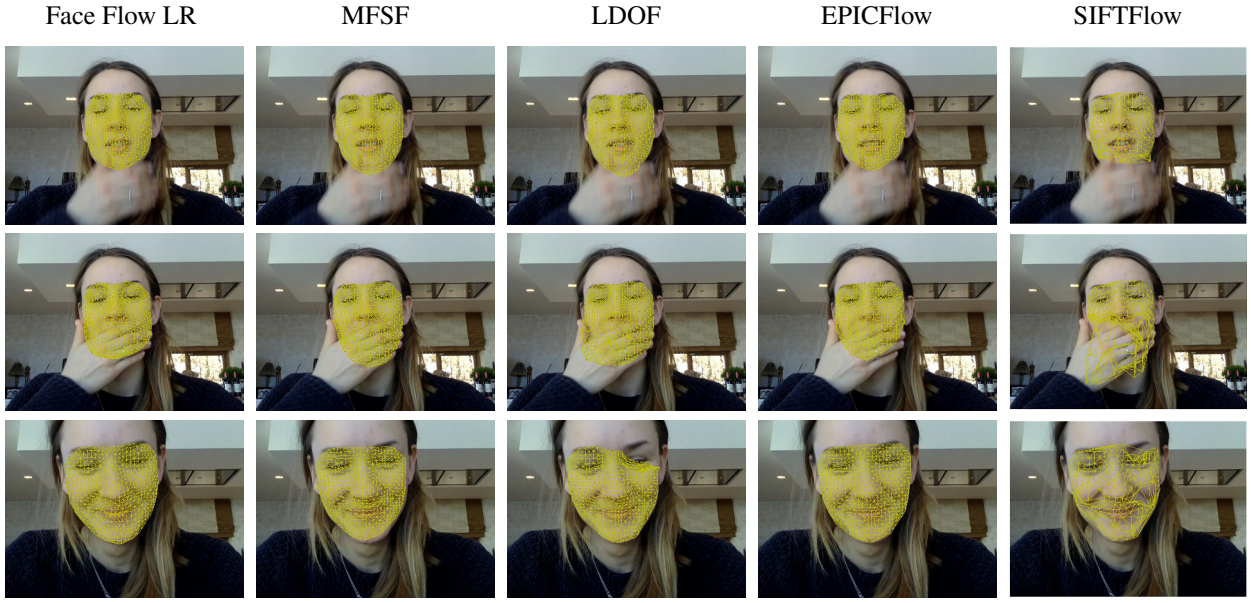

Figure 3: Results on real data. Here, we incorporate the landmark constraint and thus do not give results for Face Flow FR.

Flow to model out-of-plane rotations.

Acknowledgements

Patrick Snape is funded by a DTA from Imperial College London and by a Qualcomm Innovation Fellowship. Yannis Panagakis was funded by the ERC under the FP7 Marie Curie Intra-European Fellowship. Stefanos Zafeiriou is partially supported by the EPSRC project EP/J017787/1 (4D-FAB) and is also partially supported by the EPSRC project EP/L026813/1 Adaptive Facial Deformable Models for Tracking (ADAManT).

References

[1] J. Alabort-i-Medina, E. Antonakos, J. Booth, P. Snape, and S. Zafeiriou. Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In *ACM MM*, 2014. 4

[2] J. Alabort-i-Medina and S. Zafeiriou. Bayesian active appearance models. In *CVPR*, pages 3438–3445, 2014. 4

[3] B. Amberg, A. Blake, and T. Vetter. On compositional image alignment, with an application to active appearance models. In *CVPR*, pages 1714–1721, 2009. 5

[4] R. Anderson, B. Stenger, and R. Cipolla. Using bounded diameter minimum spanning trees to build dense active appearance models. *IJCV*, 110(1):48–57, 2013. 2

[5] E. Antonakos, J. Alabort-i-Medina, G. Tzimiropoulos, and S. Zafeiriou. Feature-based lucas-kanade and active appearance models. *IEEE TIP*, 24(9):2617–2632, 2015. 3

[6] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *CVPR*, pages 1859–1866, 2014. 4

[7] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *IJCV*, 56(3):221–255, 2004. 2, 4

[8] T. Beeler, F. Hahn, D. Bradley, B. Bickel, P. Beardsley, C. Gotsman, R. W. Sumner, and M. Gross. High-quality passive facial performance capture using anchor frames. *TOG*, 30(4):75, 2011. 2

[9] T. Brox and J. Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE T-PAMI*, 33(3):500–513, 2011. 1, 5

[10] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE T-PAMI*, 23(6):681–685, 2001. 2, 3

[11] D. Cosker, E. Krumhuber, and A. Hilton. A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In *ICCV*, pages 2296–2303, 2011. 1

[12] D. Decarlo and D. Metaxas. Optical flow constraints on deformable models with applications to face tracking. *IJCV*, 38(2):99–127, 2000. 2

[13] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *CVPR*, pages 1272–1279, 2013. 2

[14] R. Garg, A. Roussos, and L. Agapito. A variational approach to video registration with subspace constraints. *IJCV*, 104(3):286–314, 2013. 1, 2, 4, 5, 6, 7

[15] M. Irani. Multi-frame correspondence estimation using subspace constraints. *IJCV*, 48(3), 2002. 1

[16] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, pages 1867–1874, 2014. 2, 3, 4, 7

[17] I. Kemelmacher-Shlizerman and S. M. Seitz. Collection flow. In *CVPR*, pages 1792–1799, 2012. 3

[18] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE T-PAMI*, 32(11):1940–1954, 2010. 2

[19] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen. A spontaneous micro-expression database: Inducement, collection and baseline. In *FG*, pages 1–6, 2013. 2

[20] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE T-PAMI*, 33(5):978–994, 2011. 4, 5, 7

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. In *IJCV*, volume 2, pages 91–100, 2004. 4

[22] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *AI*, 1981. 2, 4

[23] Z. Luo and P. Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35, 1992. 5

[24] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, 2004. 2, 3

[25] G. Papamakarios, Y. Panagakis, and S. Zafeiriou. Generalised scalable robust principal component analysis. In *BMVC*, 2014. 1

[26] K. Ramnath, S. Baker, I. Matthews, and D. Ramanan. Increasing the density of active appearance models. In *CVPR*, pages 1–8, 2008. 2

[27] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid. EpicFlow: Edge-Preserving Interpolation of Correspondences for Optical Flow. In *CVPR*, 2015. 1, 5, 7

[28] S. Ricco and C. Tomasi. Dense lagrangian motion estimation with occlusions. In *CVPR*, pages 1800–1807, 2012. 1

[29] S. Ricco and C. Tomasi. Simultaneous compaction and factorization of sparse image motion matrices. In *ECCV*, pages 456–469. 2012. 1

[30] S. Ricco and C. Tomasi. Video motion for every visible point. In *ICCV*, pages 2464–2471, 2013. 1

[31] M. Rubinstein, C. Liu, and W. T. Freeman. Towards longer long-range motion trajectories. In *BMVC*, pages 1–11, 2012. 1

[32] P. Sand and S. Teller. Particle video: Long-range motion estimation using point trajectories. *IJCV*, 80(1):72–91, 2008. 1

[33] D. Sondermann. Best approximate solutions to matrix equations under rank restrictions. *Statistische Hefte*, 27(1):57–66, 1986. 5

[34] N. Sundaram, T. Brox, and K. Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *ECCV*, pages 438–451. 2010. 1

[35] L. Torresani and C. Bregler. Space-time tracking. In *Image Analysis and Processing*, pages 801–812. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002. 1

[36] L. Torresani, D. B. Yang, E. J. Alexander, and C. Bregler. Tracking and modeling non-rigid objects with rank constraints. In *CVPR*, pages –I–500, 2001. 1

[37] A. Wedel, T. Pock, C. Zach, H. Bischof, and D. Cremers. An improved algorithm for TV-L1 optical flow. In *Statistical and Geometrical Approaches to Visual Motion Analysis*, LNCS, pages 23–45. Springer Berlin, 2009. 2

[38] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. In *ICCV*, pages 1385–1392, 2013. 7

[39] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539, 2013. 2

[40] Y. Yacoob and L. S. Davis. Recognizing human facial expressions from long image sequences using optical flow. *IEEE T-PAMI*, 18(6):636–642, 1996. 2

[41] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *FG*, pages 211–216, 2006. 6

[42] S. Zafeiriou, C. Zhang, and Z. Zhang. A survey on face detection in the wild: Past, present and future. *CVIU*, 138:1 – 24, 2015. 4

[43] L. Zhang, N. Snavely, B. Curless, and S. M. Seitz. Spacetime faces: High-resolution capture for modeling and animation. In *Data-Driven 3D Facial Animation*, pages 248–276. Springer, 2008. 5, 6